IN THE UNITED STATES DISTRICT COURT
FOR THE MIDDLE DISTRICT OF TENNESSEE
NASHVILLE DIVISION

CONCORD MUSIC GROUP, INC., ET AL., Plaintiffs, v. ANTHROPIC PBC, Defendant.	Case No. 3:23-cv-01092 Chief Judge Waverly D. Crenshaw, Jr. Magistrate Judge Alistair Newbern

DECLARATION OF BEN Y. ZHAO

I, **Ben Y. Zhao**, hereby declare pursuant to 28 U.S.C. § 1746:

1. I am the Neubauer Professor of Computer Science at the University of Chicago, and a Fellow of the Association for Computing Machinery ("ACM").

2. I submit this declaration in connection with Plaintiffs' Reply in Support of the Motion for Preliminary Injunction filed by Plaintiffs Concord Music Group, Inc., Capitol CMG, Inc., Universal Music Corp., Songs of Universal, Inc., Universal Music – MGB NA LLC, Polygram Publishing, Inc., Universal Music – Z Tunes LLC, and ABKCO Music, Inc. (collectively, "Publishers"), in their lawsuit against Defendant Anthropic PBC ("Anthropic").

3. My statements set forth below are based on my specialized knowledge, education, and experience, as applied to the facts and circumstances of this case. If called upon, I would and could competently testify as to the matters contained herein.

4. This declaration incorporates by reference my previous declaration dated November 16, 2023, ECF No. 47, submitted in support of Plaintiffs' Motion for Preliminary Injunction, ECF Nos. 40–41 ("November 16, 2023 Declaration"). My November 16, 2023

Declaration contains important background and details not repeated here, including my curriculum vitae, areas of expertise, and initial conclusions relating to Anthropic's training of Claude with Plaintiffs' copyrighted works and the delivery of those works as outputs.

5. In preparing this declaration, I relied on my general background and training and reviewed Publishers' filings, my previous declaration, Anthropic's Opposition to Plaintiffs' Motion for Preliminary Injunction ("Anthropic's Opposition") and supporting filings, and other documents referenced herein.

6. The purpose and intent of this declaration is to provide a limited response to assertions made in Anthropic's Opposition.

Terminology

7. Anthropic's Opposition suggests that my use of the term "attack" when referring to Plaintiffs' queries submitted to Claude demonstrates that these queries were not natural or representative of how ordinary consumers would use Claude. This is not accurate. "Attack" is a term of art in computer science to describe a technique that may reveal flaws in a system, including in AI models. For example, AI security researchers Nicholas Carlini et al. use the term "attack" 18 times, including the introduction, of their 2023 paper "Quantifying Memorization Across Neural Language Models,"[1] cited in my November 16, 2023 Declaration. There is nothing nefarious or even particularly notable about this term, and my use of it certainly does not demonstrate that the prompts used by Plaintiffs were unforeseeable. Indeed, as described below, they were not only foreseeable, but anticipated by Anthropic.

[1] Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* (Mar. 6, 2023), arXiv: 2202.07646, https://arxiv.org/pdf/2202.07646.pdf.

Storage of Plaintiffs' Works

8. Anthropic asserts in its Brief in support of its Opposition that it does not "store" Plaintiffs' works.[2] I presume Anthropic means that it does not preserve a copy of the original training data as readable text. I am unable to determine to what extent that statement is accurate. But at a minimum, Anthropic must store Plaintiffs' works in tokenized format in the model, and can re-integrate them on command in response to user queries. That is how Claude is able to output complete copies of Plaintiffs' songs and how Anthropic is able to build guardrails designed to prevent the output of the specific 500 works in suit. If it did not store these works in some fashion, Anthropic would not be able to build a guardrail tailored to these particular sets of text.

Claude Training and Finetuning

9. I have reviewed the Declaration of Jared Kaplan, ECF No. 55-1 ("Kaplan Decl."), submitted in support of Anthropic's Opposition. In Paragraph 28, Mr. Kaplan states, "Anthropic has no interest in extracting or retaining the actual expressive content of any particular text used to train Claude. Claude does not retain specific quotes or chunks of texts from the corpus for the purpose of storing information it can regurgitate in response to a user prompt." Anthropic's Brief in support of its Opposition further argues that Claude's ability to output copyrighted material is a "'bug' not a 'feature' of the product."[3] These assertions are contradicted by the record of Claude's training.

10. In Paragraph 14 of his declaration, Mr. Kaplan cites to Claude's Model Card (a model card is a short statement describing the purpose of the model), attached as Exhibit A to the Kaplan Decl., as an authoritative statement of Claude's training inputs. The Claude Model Card

[2] Anthropic's Opposition at 6, 15–16.
[3] *Id.* at 2.

states that "[s]ome of the human feedback data used to finetune Claude was made public," citing to the dataset hosted on the Hugging Face website.[4]

11. That Anthropic dataset contains numerous instances where Anthropic used song lyrics to "finetune" Claude.[5] As described in my November 16, 2023 Declaration, the purpose of finetuning is to help the LLM give more precise and tailored responses, consistent with the developer's expected and intended use of the model.[6] Finetuning is typically done by individuals who either are the intended users of the model, or are trained to develop queries in the style of the intended users. Some of Anthropic's finetuning prompts that demonstrate that Anthropic used lyrics-related prompts to fine tune the model are:

- "What are the lyrics to American Pie by Don McLean?"
- "Do you know the lyrics to 'I am The Walrus'? … Can you give me the first verse?"
- "Please provide the lyrics for the song 'Only Hope' by Mandy Moore."
- "Please retype the lyrics, to the song, Mad About You, by Sting."
- "I'm trying to figure out an artist by using the lyrics. Are you able to help?"
- "I want to know the complete lyrics."
- "Can you help me identify the name of the song that includes the following lyrics[?]"

---

[4] Hugging Face is a "machine learning (ML) and data science platform and community that helps users build, deploy and train machine learning models…. Users can also browse through models and data sets that other people have uploaded." Ben Lutkevich, *What is Hugging Face?*, TECHTARGET (Sept. 2023), https://www.techtarget.com/whatis/definition/Hugging-Face#:~:text=Hugging%20Face%20is%20a%20machine,(AI)%20in%20live%20applications.

[5] *Dataset Card for HH-RLHF*, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Feb. 13, 2024).

[6] November 15, 2023 Declaration ¶ 19.

These are just a handful of examples. A more complete list can be found in the Declaration of Attorney Timothy Chung ("Chung Decl."), Ex. A. In response to these prompts, Claude was presented with two competing outputs and was trained to select the better one. Generally, the two choices are quite similar and differ primarily in the way the output is expressed. This is because the finetuning stage is not the point at which content is chosen for the model, but rather when the model is trained in preferred ways to respond in various contexts. In these examples, the context is that of song lyrics. As described in the Chung Decl.,[7] among the songs Anthropic included in its prompts, as reflected in the Hugging Face data are:

- "All Along the Watchtower" as performed by Bob Dylan
- "Brown Sugar" as performed by the Rolling Stones
- "Can't Get You Out of My Head" as performed by Kylie Minogue
- "In the Air Tonight" as performed by Phil Collins
- "My Heart Will Go On" as performed by Celine Dion
- "November Rain" as performed by Guns N' Roses
- "Roar" as performed by Katy Perry

12. The fact that Anthropic trains Claude extensively with lyrics as prompts indicates its recognition that subscribers will utilize Claude to seek lyrics to well-known songs. The finetuning data (for example, the prompt "Rewrite the lyrics to Paradise City in the style of a Dr. Seuss book") also shows that Anthropic anticipated that consumers may seek to use well-known lyrics to create new songs that incorporate recognizable portions of pre-existing works. The use of lyrics in this was part of the feature set of the Claude LLM. Finally, the finetuning data indicates

[7] Reply Declaration of Attorney Timothy Chung, Ex. A.

that the types of prompts directed to Claude by Plaintiffs are exactly the sort that ordinary users of Claude were expected to construct.

Claude's [redacted] and Guardrails

13. Mr. Kaplan's declaration describes [redacted]

14. Mr. Kaplan asserts that "Anthropic's guardrails were effective" but provides a mere two examples of [redacted] successfully blocking outputs prior to the filing of the Complaint.[10] Mr. Kaplan has not cited the overall failure rates (false positives and false negatives) of [redacted] which is a basic technical characteristic of [redacted] and should be known to Anthropic. At best, he provides two examples to show that [redacted] occasionally functioned.[11] As is evident from the fact that Plaintiffs were able to obtain verbatim or near-verbatim copies of the 500 works in suit, Anthropic's [redacted] were not effective.

15. Mr. Kaplan does not specify whether [redacted] were intentionally designed or allowed to be porous in order to make Claude a more consumer-friendly experience. But whatever Anthropic's intent, certainly we can conclude that Claude's [redacted] were not effective with respect to lyrics.

[8] Kaplan Decl. ¶ 36.
[9] *Id.* ¶ 37 (emphasis added).
[10] *Id.* ¶ 38.
[11] *Id.* ¶¶ 38–39.

16. In a clear indication that its [redacted] were not sufficient, Anthropic implemented additional guardrails after suit was filed. Mr. Kaplan does not describe how these new guardrails work or give any technical details whatsoever regarding their design, implementation, and functioning. He lists one sample interaction in which Claude refused to provide requested song lyrics, citing copyright restrictions.[12] But even before October 2023, Claude would at least occasionally activate a copyright guardrail.[13] Producing one additional example post-Complaint does not demonstrate the effectiveness of the new guardrails. Again, Mr. Kaplan fails to cite more general failure rate testing.

17. Mr. Kaplan states that Anthropic [redacted] Yet he does not explain on what technical bases are these judgments made and whether [redacted]

18. Moreover, [redacted] By definition, they will be incapable of protecting newer works, works by lesser-known creators, or anything [redacted]

19. Indeed, I understand that [redacted] continue to fall short of their promised effectiveness. Plaintiffs were able to obtain copies of many works in suit through Claude

[12] *Id.* ¶ 40.
[13] *See id.* ¶¶ 38–39.
[14] *Id.* ¶ 41.
[15] *Id.*

Instant using the Poe application[16]—a website that aggregates access to several AI models on one platform. It therefore appears that [redacted] do not in all cases penetrate to third-party applications and services using all versions of Claude. Since such third-party products and services appear to be a substantial aspect of Anthropic's intended business,[17] this problem is likely not limited to Poe and potentially implicates many users. Plaintiffs have also been able to obtain infringing outputs directly from Claude since the Complaint was filed.[18] [redacted] the Complaint was filed (and apparently as a direct result of the lawsuit), more work is necessary to prevent the infringement of copyrighted works.

20. In my prior Declaration, I noted that memorization increases with model scale and with repetition within training data.[19] Mr. Kaplan states that the next version of Claude will be trained on an even larger data set.[20] A larger training dataset creates more opportunities for repetition in outputs. He does not explain what steps have been taken to reduce duplication. The [redacted] and guardrails were insufficient with Claude 2 and had to be supplemented. Now, a new model is coming that likely has increased capacity for memorization. Anthropic has offered no evidence to provide quantitative assurance that the [redacted]

21. More generally, I believe Anthropic's argument, that it is consistently able to patch whatever holes Plaintiffs find, is fundamentally flawed. This lawsuit is not an exercise in "red

[16] Reply Declaration of Michael Candore, Ex. A.
[17] Anthropic recently announced a deal with Amazon Web Services to provide Claude "to organizations around the world." Manish Singh, *Amazon to invest up to $4 billion in AI startup Anthropic*, TECHCRUNCH (Sept. 25, 2023), https://techcrunch.com/2023/09/25/amazon-to-invest-up-to-4-billion-in-ai-startup-anthropic/.
[18] Reply Declaration of Attorney Timothy Chung, Ex. B.
[19] November 16, 2023 Declaration ¶ 24 n.9.
[20] Kaplan Decl. ¶ 23.

teaming," where Plaintiffs identify vulnerabilities in Claude so Anthropic can fix the system it built. And to the extent Plaintiffs do share the results of their testing, they are not responsible for identifying all the flaws in the Claude system. Anthropic should be proactively working to prevent the Publishers' lyrics from being incorporated into its training data and outputs. Nothing in Anthropic's Opposition indicates that it has taken a more forward-looking, aggressive approach to addressing this problem, which after all, is one of its own making.

Cost to Remove Plaintiffs Works from Training Dataset

22. Mr. Kaplan states [redacted]

I doubt the accuracy of these estimates. The vast majority of the cost of building an AI model is in the hardware and software needed to run the system, and in the iterative design of the training pipeline that requires repeated testing and evaluation of [redacted] optimization algorithms, storage/access components by engineers. However, these are one-time costs that Anthropic has already spent. Training the model from scratch takes computation time, but removing a small portion of the dataset should only involve pure computation of the curated training tokens through the established pipeline, and should not involve additional costly engineering or testing, which should make up the bulk of the costs in updating a model. I believe Anthropic should be able to [redacted]

[21] *Id*. ¶ 44.

I declare under penalty of perjury under the laws of the United States that the foregoing is true and correct to the best of my personal knowledge and belief.

Executed in Chicago, IL, this 14th day of February, 2024.

Ben Y. Zhao

CERTIFICATE OF SERVICE

I hereby certify that on February 15, 2024, I authorized the electronic filing of a true and exact copy of the foregoing with the Clerk of the Court using the CM/ECF system, which sent notice of such filing to the following:

Aubrey B. Harwell III (No. 017394)
Nathan C. Sanders (No. 33520)
Olivia R. Arboneaux (No. 40225)
NEAL & HARWELL, PLC
1201 Demonbreun Street, Suite 1000
Nashville, TN 37203
tharwell@nealharwell.com
nsanders@nealharwell.com
oarboneaux@nealharwell.com

Joseph R. Wetzel
Andrew M. Gass
LATHAM & WATKINS LLP
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
joe.wetzel@lw.com
andrew.gass@lw.com

Allison L. Stillman
LATHAM & WATKINS LLP
1271 Avenue of the Americas
New York, NY 10020
alli.stillman@lw.com

Sarang V. Damle
LATHAM & WATKINS LLP
555 Eleventh Street, NW, Suite 1000
Washington, DC 20004
sy.damle@lw.com

Kevin C. Klein
KLEIN SOLOMON MILLS, PLLC
1322 4th Avenue North
Nashville, TN 37208
kevin.klein@kleinpllc.com

Eric P. Tuttle
WILSON SONSINI GOODRICH & ROSATI
701 Fifth Avenue, Suite 5100
Seattle, WA 98104
eric.tuttle@wsgr.com

Nicole Saad Bembridge
NETCHOICE, LLC
1401 K St. NW, Suite 502
Washington, DC 20005
nsaadbembridge@netchoice.org

Jacob T. Clabo
Lauren E. Kilgore
SHACKELFORD BOWEN MCKINLEY NORTON, LLP
1 Music Circle South
Suite 300
Nashville, TN 37203
jclabo@shackelford.law
lkilgore@shackelford.law

Frank P. Scibilia
Maya B. Katalan
PRYOR CASHMAN LLP
7 Times Square, 40th Floor
New York, NY 10036
fscibilia@pryorcashman.com
mkatalan@pryorcashman.com

/s/ *Timothy Chung*
Timothy Chung